

A Phonotactic-Semantic Paradigm for Automatic Spoken Document Classification

Bin MA, Haizhou LI
Institute for Infocomm Research
21 Heng Mui Keng Terrace, Singapore 119613
{mabin, hli}@i2r.a-star.edu.sg

ABSTRACT

We demonstrate a *phonotactic-semantic* paradigm for spoken document categorization. In this framework, we define a set of acoustic words instead of lexical words to represent acoustic activities in spoken languages. The strategy for acoustic vocabulary selection is studied by comparing different feature selection methods. With an appropriate acoustic vocabulary, a voice tokenizer converts a spoken document into a text-like document of acoustic words. Thus, a spoken document can be represented by a count vector, named a *bag-of-sounds* vector, which characterizes a spoken document's semantic domain. We study two *phonotactic-semantic* classifiers, the support vector machine classifier and the latent semantic analysis classifier, and their properties. The *phonotactic-semantic* framework constitutes a new paradigm in spoken document classification, as demonstrated by its success in the spoken language identification task. It achieves 18.2% error reduction over state-of-the-art benchmark performance on the 1996 NIST Language Recognition Evaluation database.

Categories and Subject Descriptors

I.5.2 [Pattern Recognition]: Design Methodology – Classifier design and evaluation; H.5.5 [Information Interfaces and Presentation]: Sound and Music Computing – Methodologies and techniques.

General Terms

Algorithms, Performance, Experimentation, Languages.

Keywords

Spoken Document Classification, Voice Tokenizer, Acoustic Words, Semantic Domain, *n*-gram, Phonotactic-Semantic.

1. INTRODUCTION

With the rapid expansion of audio media sources such as radio, TV and telephony recordings, there is an increasing demand for

automatic indexing and retrieval of spoken documents. Spoken Document Retrieval (SDR) is essentially the task of retrieving excerpts from a large collection of spoken documents based on a user's request. This task has been conventionally approached by integrating text information retrieval (IR) and automatic speech recognition (ASR) technologies. Due to the constant efforts in IR and ASR, recent research in SDR has achieved major advances. Research in SDR has reached such a level of interest that the Text REtrieval Conference (TREC) has even included an SDR track from 1997 to 2000 [7].

Automatic spoken document classification (SDC) is an important topic in SDR, and has been given much attention. Most SDC efforts so far have been devoted to the *lexical-semantic* and the *n-gram phonotactic* paradigms. In the *lexical-semantic* paradigm, text categorization (TC) techniques are applied to the automatic transcripts of spoken documents to derive semantic classes. The transcripts are typically generated from a large vocabulary continuous speech recognizer (LVCSR). In a nutshell, the *lexical-semantic* method simply cascades a LVCSR and a TC module. In the *n-gram phonotactic* paradigm, the idea is to use *n*-gram phonotactics, i.e. the rules governing the sequences of allowable phonemes, instead of lexical words to represent the lexical constraints that are imposed by semantic domains, in an effort to enhance robustness against speech recognition errors.

However, the task of SDC is more complex than the TC task. By comparing them, we can gain some insights into the SDC task and the inadequacy of the TC methods that have been applied to SDC. In TC, we usually derive the lexical vocabulary from the running text. However, for spoken documents, an additional tokenization step is needed to convert sound wave into a sequence of phonetic units, such as phonemes. This gives rise to two issues: the definition of tokenization unit, and the choice of vocabulary. These two issues have direct impacts on the resulting tokenization and the subsequent SDC performance. To address them, let's look into two intrinsic properties of spoken language.

First, to properly select a vocabulary, we need to take into account Zipf's Law[17]. In human languages, some words invariably occur more frequently than others. One of the most common ways of expressing this idea is known as Zipf's Law. This law states that there is always a set of words which dominates most of the other words of the language in terms of their frequency of use. This is true both of words in the general domain and of words that are specific to a particular subject or semantic domain. This is also true both of written words and spoken words. In SDC, we are particularly interested in extracting a vocabulary that is semantically discriminative.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SIGIR '05, August 15-19, 2005, Salvador, Brazil.

Copyright 2005 ACM 1-59593-034-5/05/0008...\$5.00.

Second, in automatic spoken document classification (SDC), the tokenization unit has traditionally been the lexical word. However, since the lexical word is just the written convention of the language, there is no strong reason why we should choose it as the tokenization unit. We propose to use language independent *acoustic word* (AW) as an alternative. A lexical word is usually defined by its semantic and/or syntactic function while an AW is associated with a sequence of sounds, for instance, phoneme n -gram. With this bold proposal, we look forward to deriving semantic classes of spoken documents based on AW statistics, also referred to as *bag-of-sounds*.

This paper is organized as follows. In Section 2, through discussion of related work, we highlight the areas of interest. In Section 3, we propose the *bag-of-sounds* representation of spoken documents. In Section 4, we formulate SDC under the *phonotactic-semantic* framework using both support vector machines (SVM) and latent semantic analysis (LSA) techniques. We evaluate the proposed *phonotactic-semantic* approach on one of the SDC tasks, spoken language identification. In Section 5, we conclude our study and discuss future work.

2. RELATED WORK

Spoken documents, in the form of digitized wave files, are far less structured than written documents and need to be treated with techniques that go beyond the bounds of written language. The challenge of SDC is inter-disciplinary, involving digital signal processing, speech recognition and natural language processing. In general, a SDC system has three fundamental components:

- 1) A voice tokenizer, i.e. a speech recognizer front-end which segments a spoken documents into acoustic tokens;
- 2) A statistical language model which captures statistics of semantic domain information;
- 3) A classifier which categorizes a spoken document using the statistical language model.

Prior work in this area can be separated into of two categories: the *lexical-semantic* approach and the *n-gram phonotactic* approach.

2.1 The *Lexical-semantic* Approach

The *lexical-semantic* approach originates from text-categorization. The idea is to convert the spoken documents into text transcripts of lexical words, then categorize the transcripts as if they were text documents. Research has revealed that machine learning techniques such as SVM and LSA are effective in handling high dimension feature vector classification and are robust to speech recognition error as far as SDC is concerned [4,5]. SVM and LSA are used to induce high level latent semantic classes from the sparsely populated data space. This approach has proven effective under a few assumptions: first, the speech recognizer uses a large vocabulary containing most of the words to be recognized; second, the spoken document is consistent with the language model used. Despite many successes, the *lexical-semantic* approach suffers from several shortcomings because its assumptions don't always hold in real-world applications.

- 1) **Homophone:** Due to imprecision of automatic transcripts, homophones, or rather acoustically confusable lexical words, degrade the semantic clustering performance. The example "to recognize speech" vs. "to wreck a nice beach" illustrates

totally different semantic induction. A lexical-based language model built from the first transcript does not work for the second transcript at all.

- 2) **Out-of-Vocabulary (OOV):** Unlike text documents where lexical words can be extracted from the running text, spoken documents may include many new words that a pre-defined vocabulary for automatic speech recognition (ASR) does not cover. As a result, the SDC performance is especially compromised when OOV problem becomes acute, e.g. in spoken collection of names and places.
- 3) **Multilinguality:** Acoustic vocabularies can be language independent, whereas lexical vocabularies are language dependent. The *lexical-semantic* approach uses lexical word as tokenization unit that makes it difficult to accommodate multiple languages. As a result, *lexical-semantic* approach has been confined to monolingual SDC so far.

To overcome the homophone and OOV problems, efforts have been made to reduce the affection of errors from ASR by representing a group of acoustically similar lexical words with an *acoustic word* (AW), such as soundex word [5] or subword [12,13]. These solutions move from lexical-based towards acoustic-based semantic induction for SDC. However, the reported methods derive AWs from a language dependent phonetic lexicon instead of speech transcripts, thus inheriting the limitations of the *lexical-semantic* framework. The multilinguality problem can't possibly be resolved using the *lexical-semantic* approach.

2.2 The *n-gram Phonotactic* Approach

Despite many differences, spoken language and written language are also similar in many ways. For example, both text and voice can be seen as stochastic time-sequences corrupted by a channel noise. The *n-gram* language model has achieved equal amounts of success in both tasks, e.g. n -character slice modeling for text categorization by language [3], and Phone Recognition followed by n -gram Language Modeling (PRLM) for spoken language identification (SLID) [18].

A successful experiment was reported in [1] for automatic call routing, the task of categorizing a customer's telephone enquiry and automatically relaying it to one of its appropriate destinations based on its semantic content. Using an n -gram language model, named the *n-gram phonotactic* approach, the SDC performance achieved was surprisingly close to that of the *lexical-semantic* approach, with the advantage of using label-free training data and reducing computing cost.

Efforts in SLID are represented by Zissman [18]. Results show that the *n-gram* language model effectively captures the lexical constraints that are imposed by a spoken language. Formal evaluations of SLID conducted by the National Institute of Science and Technology (NIST) in recent years demonstrated that the *n-gram phonotactic* approach is the most successful approach for the SLID problem [15, 16].

The *n-gram phonotactic approach* typically uses phoneme as the voice tokenization unit in an effort to overcome the problems that the *lexical-semantic* approach faces. It also simplifies the voice tokenizer front-end from a large vocabulary continuous speech recognizer (LVCSR) in the *lexical-semantic* framework to a phoneme recognizer. However, in the prior work, the *n-gram*

phonotactic approach leaves two outstanding problems unattended, which limits the deployment of the technology:

- 1) **Semantic Abstraction:** The *n*-gram *phonotactic* approach only captures *n*-local phonotactics to reflect lexical constraints. The semantic characteristics at the level of utterance or spoken document remain unexplored.
- 2) **Multilinguality:** Thousands of spoken languages from all over the world are phonetically articulated using only a few hundred distinctive sounds [8]. Therefore, a unified language independent phoneme set is practically feasible. However, it has not been studied experimentally so far.

2.3 Our Phonotactic-semantic Approach

Obviously, the *n*-gram *phonotactic* approach suffers the major shortcoming of not exploiting the global phonotactics in the larger context of a spoken document, while the *lexical-semantic* approach inherits limitations from its lexical choice. Their differences lie in two aspects: the lexical constraint definition and the means for latent semantic abstraction. We approach the two problems from a new perspective with the *phonotactic-semantic* paradigm:

1) By adopting language independent *acoustic word* (AW) instead of lexical word, the acoustic vocabulary can be learned from a multilingual training corpus of transcripts using a data driven approach. The novelty here lies in the learnable and language independent acoustic vocabulary.

2) A domain specific spoken document always contain a set of high frequency content words, prefixes, and suffixes, which are realized as phoneme substrings. Individually, these substrings may be shared across documents in different semantic domains. However, the pattern of their co-occurrences discriminates one semantic class from another. The novelty here is to use the *bag-of-sounds* statistics over AWs, instead of *bag-of-words* over lexical words, to derive high level semantic characteristics from a spoken document.

The comparison of the three approaches is summarized in Table 1:

Table 1. Summary of three approaches

	Lexical constraint	Latent semantics	Outstanding problems
Lexical-semantic approach	Lexical word	<i>bag-of-words</i> vector	1.Homophone 2.OOV 3.Multilinguality
<i>n</i> -gram phonotactic approach	<i>n</i> -local phonotactics		1.Multilinguality 2.Semantic Abstraction
Phonotactic-semantic approach	<i>n</i> -local phonotactics	<i>bag-of-sounds</i> vector	

3. BAG-OF-SOUNDS REPRESENTATION

Without loss of generality, we study the spoken language identification (SLID) application with our proposed paradigm in this paper. However, the new paradigm is readily applied to SDC in general.

A variety of cues can be used by humans and machines to distinguish one language from another. These cues include

phonology, prosody, morphology and syntax. The phonological and prosodic features are reflected in acoustic tokenization while morphological/lexical and syntactic features are typically described by the *n*-gram language model. The reported successful *lexical-semantic* and *n*-gram *phonotactic* approaches so far take advantage of one or more of these sets of language traits.

The *bag-of-sounds* concept is analogous to the *bag-of-words* paradigm originally formulated in the context of information retrieval (IR) and text categorization (TC) [2,4,14]. One focus of TC is to extract informative features for document representation in a vector space. It is believed that it is not just the words, but also the co-occurrence of words, that distinguish the semantic domains of text documents. In practice, a document is represented by a high-dimensional vector derived from the statistics of term frequency. To improve the vector's expressiveness and reduce its number of dimensions, algorithms such as LSA have been proposed to characterize salient semantic information across the entire text [2].

In SLID, it is similarly believed that although the sounds of different spoken languages overlap considerably, their phonotactics differentiate one language from another. Therefore, one can easily draw the analogy between an AW in *bag-of-sounds* and a lexical word in *bag-of-words*. Unlike words in a text document, the phonotactic information that distinguishes spoken languages is concealed in the sound waves of spoken languages. A step is needed to transcribe a spoken document into a text-like document of sound tokens, named phonetic transcript. Many TC techniques can then be readily applied. By assuming that the overall sound characteristics of all spoken languages can be covered by a universal collection of acoustic segment models without imposing any phonetic definitions, we adopt the language-independent phoneme-like unit as an acoustic token in the voice tokenizer [9]. A token *n*-gram forms an AW. A collection of AW forms an AW vocabulary.

3.1 Acoustic Word (AW)

The acoustic word is to reflect the short-term lexical constraint imposed by the language. For an acoustic system of *T* acoustic tokens, we potentially have $W = T^n$ AWs in the vocabulary. It was reported that there is little advantage to using $n > 2$ in the *n*-gram for PRLM [18]. In the interest of tractability, we will only use the bigram, that is $n=2$, also called the token pair, as the AW of our vocabulary. The advantages of such AW are: they are language independent; An AW is typically smaller than a lexical word, and an AW vocabulary has fewer entries than a lexical vocabulary. Therefore, AW is the most robust against speech recognition errors [5].

Instead of deriving AWs from a language dependent phonetic lexicon [1], we adopt a data driven approach. We extract all token pairs from the phonetic transcripts of spoken documents. By collecting all the token pairs that are seen in the phonetic transcripts, in the same way that we derive lexical vocabulary from text documents, we obtain an AW vocabulary. Due to the imprecision of speech recognition, not all the resulting AWs are valid. We expect to observe more AWs in the phonetic transcripts than the actual ones. Therefore, it is desirable to remove those noise AWs to reduce the feature space. One simple solution is to discard AWs that have very low frequency and AWs that occur in

too few documents to be statistically significant. This method is referred to as count-trimming (CT).

A *bag-of-sounds* vector captures document-level long-term phonotactics such as co-occurrences of AWs. We describe a spoken document as a count vector of AWs $c = \{c_1, c_2, \dots, c_W\}^T$, which has its element to represent the count of an AW and takes the AW vocabulary size W as its dimension. It is possible to explore the relations and higher-order statistics among the diverse AWs through SVM or LSA. Applying Zipf's Law, some AWs are more informative and discriminative than others. Like in text categorization, we would like to compile a list of stop AWs which do not render much discriminative information across spoken document categories. By increasing the number of stop AWs, we effectively reduce the vector dimension and subsequent computation in the classification. This is referred to as the feature selection process. To further examine the discriminative property of each feature dimension, we compare two feature selection methods based on language-labeled spoken document transcripts.

3.2 Mutual Information (MI)

Let us consider a two class situation. The class membership $X = \{x_+, x_-\}$ and a particular AW's presence $Y = \{y_+, y_-\}$ are random variables, where x_+ indicates the document is a positive sample and x_- otherwise, and y_+ indicates AW is present in a document and y_- otherwise. The mutual information as defined in information theory presents the information gain about X by knowing the presence or absence of an AW Y as:

$$MI(X, Y) = \sum_{x \in X, y \in Y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \quad (1)$$

In a multi-class situation, we would like to examine the mutual information collectively between the AW's presence Y and all the class pairs X present in the training data. The mutual information (MI) indicates how significantly an AW's presence contributes to the semantic classification of the spoken documents.

3.3 Separation Margin (SM)

SVM with a linear kernel partitions the feature space using a hyperplane $f(c) = a^T c + b$ while $a = \{a_1, a_2, \dots, a_W\}$ is the normal vector of the hyperplane determined by the classifier to separate x_+ from x_- in a 2-class situation. The hyperplane is obtained as the result of the SVM training process under a criterion to maximize the margin, or rather, to minimize the cost of misclassification. The margin is defined as the minimal distance of a sample to the decision surface. Thus, a feature c_j with the weight a_j indicates the effect of the j^{th} dimension in constructing the hyperplane. The idea is to consider the feature important if it significantly influences the width of the margin of the resulting hyperplane. This margin is inversely proportional to $\|a\|$, the length of a [11]. It was found that the features with higher $|a_j|$ are more influential in determining the width of the separation margin. A theoretical justification for retaining the

highest weighted features in the normal can be found in [10]. We also adopt this strategy in AW selection and refer it to separation margin (SM) criterion.

3.4 Feature Weighting

AW is characterized by sounds instead of part-of-speech. We can't identify stop AWs in the same way as we extract stop words in lexical vocabulary. The feature selection allows us to decide the dimension of the *bag-of-sounds* vector using data driven approach. It is equally important to weight the raw counts to refine the contribution of each AW. We begin by normalizing the vectors representing the acoustic words by making each vector of unit length. Our second weighting is based on the notion that an AW that only occurs in a few documents is more discriminative than an AW that occurs in nearly every document. We use the *inverse-document frequency (idf)* weighting scheme [4], in which an AW is weighted inversely to the number of documents in which it occurs, by means of $idf(w) = \log D / d(w)$. In this equation, w is an AW in a vocabulary of W AWs, D is the total number of documents in the corpus and $d(w)$ is the number of document containing the AW w . Let $c_{w,d}$ be the count of AW w in document d . We have the weighted count as

$$c'_{w,d} = c_{w,d} \times idf(w) / \left(\sum_{1 \leq w' \leq W} c_{w',d}^2 \right)^{1/2} \quad (2)$$

and a vector $c'_d = \{c'_{1,d}, c'_{2,d}, \dots, c'_{W,d}\}^T$ to present document d .

4. DESIGN OF CLASSIFIERS

With the *bag-of-sounds* vector, the language identification task becomes a vector classification problem. Many effective classifiers exist in machine learning for high-dimension vector classification. Most of them are of great discriminative ability. These include SVM [5,11], LSA paradigm [2,4] and neural networks. Many studies also reveal that dimensionality reduction is effective in improving expressiveness of input vectors. In this section, we formulate the *phonotactic-semantic* paradigm by designing two classifiers, the SVM classifier and LSA classifier.

The SVM classifier is trained by minimizing misclassification errors. The objective in this study is to examine the coupling behavior between feature selection and SVM learning method, which enables us to make an informed conjecture in future classifier design. On the other hand, the LSA classifier is motivated by the belief that, with *bag-of-sounds* vectors, the semantic space of spoken documents can be well partitioned through latent semantic analysis in a lower dimensional space. In contrast to SVM, the LSA classifier is trained by maximum likelihood criterion which can be formulated by the well-established *Expectation-Maximization (EM)* algorithm. The objective in this study is to examine how model size affects the spoken language identification (SLID) performance.

4.1 SVM Classifier

4.1.1 Multi-class SVM Classifier (SVMC)

The SVM is a classifier of natural choice because the attribute vectors based on *bag-of-sounds* are high dimensional and sparse. SVM allows us to partition the categories of *bag-of-sounds* vectors in a high dimensional space. It is known that SVM is a 2-class $X = \{x_+, x_-\}$ classifier. The SVM algorithm trains a classifier

of the form $f(c) = a^T \psi(c) + b$, described by a hyperplane's normal vector a and an offset b . Learning is posed as an optimization problem with the goal of maximizing the margin, i.e., the distance between the separating hyperplane $a^T \psi(c) + b = 0$ and the nearest training vectors, or rather, minimizing the classification error, such that, if $f(c) > 0$, then $c \in x_+$ and if $f(c) \leq 0$, then $c \in x_-$. An extension of this formulation also allows for a wider margin at the cost of misclassifying some of the training examples. The form of this optimization task is a quadratic programming problem and can be solved numerically. We used the SVM^{light} V6.01 program¹ to train the SVM models. This program allows us to explore both linear and non-linear SVM kernel. We work with a linear kernel SVM, that has $\psi(c) = c$, because the existing literature on text categorization (TC) indicates that the non-linear versions of these algorithm gain very little in terms of performance [10].

Let's first describe a multi-class classification strategy. For L classes, we build $L \times L / 2$ 2-class classifiers. A spoken document of unknown class goes through $L \times L / 2$ 2-class classification trials. The class that gains most of the winning votes takes all. With a selected *acoustic word* (AW) vocabulary, we proceed to study the effects of feature selection and SVM training corpus size.

4.1.2 Database

This section will experimentally analyze the performance of the proposed *phonotactic-semantic* approach using the 1996 NIST Language Recognition Evaluation (LRE) database². The database was intended to establish a baseline of performance capability for language recognition of conversational telephone speech. The database contains recorded speech of 12 languages: Arabic, English, Farsi, French, German, Hindi, Japanese, Korean, Mandarin, Spanish, Tamil, and Vietnamese. We use the training set and development set of the LDC *CallFriend* corpus³ as the training data. Each conversation in the training data is segmented into overlapping sessions of about 30 seconds each, resulting in about 12,000 sessions for each language. The evaluation set consists of 1,492 30-sec sessions, each distributed among the various languages of interest. We treat a 30-sec session as a spoken document in both training and testing. We report error rates of the 1492 test trials.

4.1.3 Feature Selection

In the experiments, we compare three feature selection methods, namely CT, MI, and SM as proposed in Section 3, by using a SVM classifier to evaluate the expressiveness of the resulting AW vocabulary in a quantitative way. We adopt 128 language independent phonemes as the acoustic tokens, which are selected from the 12 languages [9]. Since $T=128$, we have $W=16,384$ AWs

in the vocabulary. The objective of feature selection is to reduce the AW vocabulary size and thus improve the expressiveness of *bag-of-sounds* vectors. By removing less influential AWs, we gradually reduce the dimension to examine the spoken language identification (SLID) performance using a SVM with a linear kernel. The results are reported in Figure 1. As expected, the simple count-trimming technique (CT) does not work as well as discriminative selection methods. Between MI and SM, SM works slightly better. We speculate that this is probably due to the consistency between feature selection and decision strategy as SM uses SVM normal as the feature selection criterion. A similar finding was also reported in [10]. The feature selection is rather effective. We maintain almost the same error rate by reducing vocabulary size from 16,384 to 9,000, which amounts to a 45% dimension reduction.

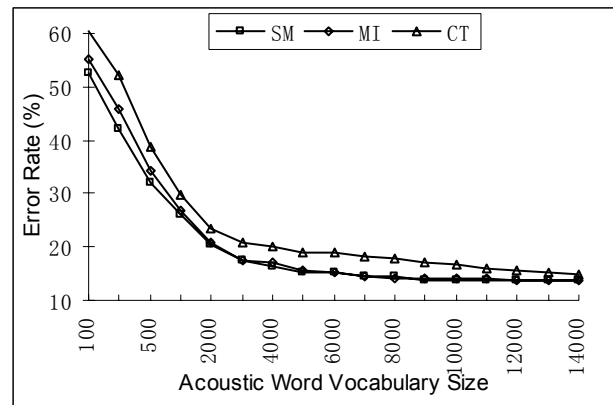


Figure 1. SLID error rate comparison among three feature selection techniques

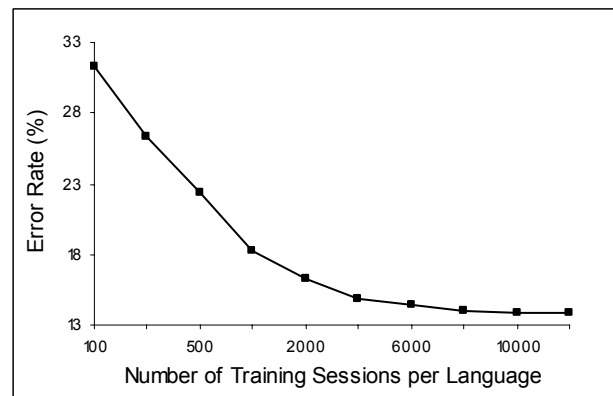


Figure 2. Effect of training corpus size

4.1.4 Effects of Training Data Size

In a real-world application, we always face the problem of insufficient training data because it is costly to collect a lot of domain-labeled samples. In our case, the domain refers to a spoken language. It is of interest to know how the amount of training data affects the performance of the resulting SVM classifier. The chart in Figure 2 reports the error rates presented by SVMs that are trained with different amount of data. The full corpus includes 12,000 spoken documents for each language. The subset corpus is randomly selected from the full corpus with equal

¹ <http://svmlight.joachims.org/>

² <http://www.nist.gov/speech/tests/index.htm>

³ See <http://www ldc.upenn.edu/>. The overlap between 1996 NIST evaluation data and *CallFriend* database has been removed from training data as suggested in the 2003 NIST LRE website <http://www.nist.gov/speech/tests/index.htm>

amount from each language. We observed that when subset corpus grows beyond 8,000 per language, the performance of SVM begins to saturate.

4.2 LSA Classifier

4.2.1 Latent Semantic Analysis

Let's start with a standard latent semantic analysis (LSA). A corpus of D documents can be represented by a term-document matrix $H : W \times D$. We aim to extract the latent phonotactic information that is defused in a given spoken document. LSA is used to decompose the matrix H into a multiplication of three matrices through Singular Vector Decomposition (SVD):

$$H = USV^T \quad (3)$$

U is a $W \times R$ left singular matrix with rows $u_w, 1 \leq w \leq W$; S is a $R \times R$ diagonal matrix of singular values $s_1 \geq s_2 \geq \dots \geq s_R > 0$; V is $D \times R$ right singular matrix with rows $v_d, 1 \leq d \leq D$.

Both the left and the right singular matrices are column-orthonormal. If we retain only the top Q singular values in matrix S and zero out the other ($R-Q$) components, the LSA feature dimension can be effectively reduced to Q , which is often much smaller than R . We can therefore compare spoken documents in this new Q dimensional space, referred to as Q -space in the rest of this paper. After SVD, it is easy to arrive at a natural metric for the closeness between two spoken documents c_i and c_j :

$$g(c_i, c_j) \approx \cos(\tilde{v}_i, \tilde{v}_j) = \frac{v_i S^2 v_j^T}{\|v_i S\| \cdot \|v_j S\|} \quad (4)$$

The similarity between two vectors, $g(c_i, c_j)$, is approximated by the cosine of the angle of $\tilde{v}_i = v_i S$ and $\tilde{v}_j = v_j S$, where the tilde indicates vector in the Q -space. It can be transformed to a distance measure $k(c_i, c_j) \approx k(\tilde{v}_i, \tilde{v}_j) = \cos^{-1} g(c_i, c_j)$.

In forced-choice SLID, a test spoken document c_p , which is not part of the training data, is classified into one of the L languages. We assume consistency between the test document's intrinsic phonotactic pattern and one of the patterns that is extracted from the L languages so that the SVD matrices still apply to the test document. Using the left singular matrix U , one is able to construct a document vector \tilde{v}_p in the Q -space, referred to as the *pseudo document vector* [4],

$$c_p \rightarrow \tilde{v}_p = c_p^T U S^{-1} \quad (5)$$

The LSA framework leads to a number of interesting properties in the Q -space. It allows us to explore various classifiers under the *minimum distance* and the *maximum likelihood* criteria.

4.2.2 LSA Classifier I

The pattern of language distribution is inherently multi-modal, so it is unlikely to be well fitted by a single vector. Suppose that we have M prototypical vectors, also referred to as the centroids, to represent a language. Applying LSA to a term-document matrix $H : W \times L'$, where $L' = L \times M$ assuming each language l is represented by a set of M vectors, Φ_l , a *minimum distance*

classifier can be formulated in the Q -space. Applying k -nearest neighboring rule [6], we have LSA classifier I (LSAC-I) as:

$$\hat{l} = \arg \min_i \sum_{l' \in \phi_i} k(\tilde{v}_p, \tilde{v}_{l'}) \quad (6)$$

where ϕ_i is the set of k -nearest-neighbors to \tilde{v}_p and $\phi_i \subset \Phi_l$. There are different ways to derive the M vectors for each language. Suppose that we have a set of training documents D_l for language l . We choose to carry out vector quantization (VQ) to partition D_l into M cells $D_{l,m}$ in the Q -space such that $\cup_{m=1}^M D_{l,m} = D_l$. Therefore, all the documents in each cell $D_{l,m}$ can be merged to form a super-document, which is then projected into a Q -space vector $\tilde{v}_{l,m}$. This results in M prototypical centroids $\tilde{v}_{l,m} \in \Phi_l$ ($m = 1, \dots, M$). Using LSAC-I, a test vector is compared with M vectors to arrive at the k -nearest neighbors for each language. This is computationally expensive, especially when M is large.

4.2.3 LSA Classifier II

Alternatively, one can account for multi-modal distribution through finite mixture model. The mixture model represents the M discrete components with soft combination. To extend the LSAC into a statistical framework, it is necessary to map our distance measure $k(\tilde{v}_i, \tilde{v}_j)$ into a probability measure. One way is for the distance measure to induce a family of exponential distributions with pertinent marginality constraints. In practice, what we need is a reasonable probability distribution, which sums to one, to act as a lookup table for the distance measure. Here we choose to use the empirical multivariate distribution constructed by allocating the total probability mass in proportion to the distances observed with the training data. In short, this reduces the task to a *histogram normalization*. In this way, we map the distance $k(\tilde{v}_i, \tilde{v}_j)$ to a conditional probability distribution $p(\tilde{v}_i | \tilde{v}_j)$

subject to $\sum_{i=1}^D p(\tilde{v}_i | \tilde{v}_j) = 1$. Now that we are in the probability

domain, techniques such as mixture smoothing can be readily applied to model a language class with finer fitting.

Let's re-visit the task of L language forced-choice classification. Denote by $\lambda_l \in \Lambda$ ($l = 1, \dots, L$) the *phonotactic-semantic* model for category l . Suppose that we have M centroids $\tilde{v}_{l,m} \in \Phi_l$ ($m = 1, \dots, M$) in the Q -space for each language l where each centroid represents a class. The class conditional probability can be described as a linear combination of $p(\tilde{v}_i | \tilde{v}_{l,m})$:

$$p(\tilde{v}_i | \lambda_l) = \sum_{m=1}^M p(\tilde{v}_{l,m}) p(\tilde{v}_i | \tilde{v}_{l,m}) \quad (7)$$

the probability $p(\tilde{v}_{l,m})$ functionally serves as a mixture weight of $p(\tilde{v}_i | \tilde{v}_{l,m})$.

A mixture model λ_l is represented by a set of centroids $\tilde{v}_{l,m} \in \Phi_l$ ($m = 1, \dots, M$) with the probability distribution of $p(\tilde{v}_i | \tilde{v}_{l,m})$ and $p(\tilde{v}_{l,m})$. $p(\tilde{v}_i | \tilde{v}_{l,m})$ is estimated by *histogram*

normalization and $p(\tilde{v}_{l,m})$ is estimated under the maximum likelihood criteria, $p(\tilde{v}_{l,m}) = C_{m,l} / C_l$, where C_l is the total number of documents in D_l , of which $C_{m,l}$ documents fall into the cell m . An iterative process of EM algorithm is devised for model training to maximize the likelihood as in Eq.(8), given a training set Ω , $\cup_{l=1}^L D_l = \Omega$. $p(\tilde{v}_i | \tilde{v}_{l,m})$ is used for M clustering instead of VQ.

$$p(\Omega | \Lambda) = \prod_{l=1}^L \prod_{d=1}^{|D_l|} p(\tilde{v}_d | \lambda_l) \quad (8)$$

A maximum-likelihood classifier, named LSA classifier II (LSAC-II) can be formulated as follows:

$$\hat{l} = \arg \max_l p(\tilde{v}_p | \lambda_l) = \arg \max_l \sum_{m=1}^M p(\tilde{v}_{l,m}) p(\tilde{v}_p | \tilde{v}_{l,m}) \quad (9)$$

Expectation-Maximization algorithm

Step 1: For each l , partition D_l into M cells $D_{l,m}$ in Q -space, such that $\cup_{m=1}^M D_{l,m} = D_l$, to initialize Λ ;

Step 2: Computing $p(\tilde{v}_i | \tilde{v}_{l,m})$ and $p(\tilde{v}_{l,m})$ for each l ;

Step 3: For each l , derive M centroids $\tilde{v}_{l,m} \in \Phi_l$ for Λ' ;

Step 4: Set $\Lambda = \Lambda'$, repeat from Step 2 until convergence.

4.2.4 Experiments

We continue to use the short-listed $W=9,000$ vocabulary resulting from Section 4.1.3, and $Q = 90$ for the LSAC experiments.

Distance Ratio

First of all, let us examine how LSA improves semantic clustering of bag-of-sounds vectors. The inter- and intra- language distance ratio indicates the data separation in the semantic space.

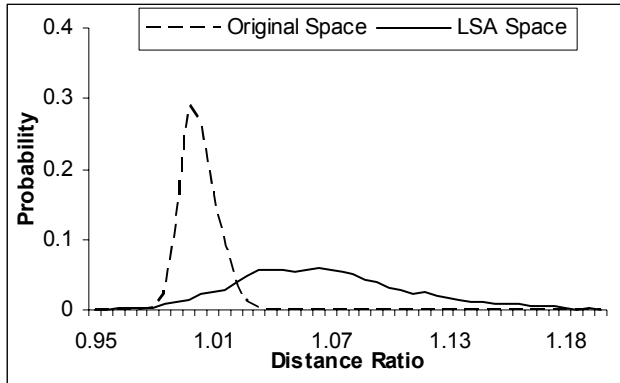


Figure 3. r_r and r_q distribution

Figure 3 illustrates the distribution of inter- and intra- language distance ratio for 3,600 randomly selected spoken documents $r_r = k(c_i, c_{inter}) / k(c_i, c_{intra})$ in the original space and $r_q = k(\tilde{v}_i, \tilde{v}_{inter}) / k(\tilde{v}_i, \tilde{v}_{intra})$ in its corresponding Q -space. We

observed that, r_q is improved over r_r . This shows LSA not only reduces the computational dimensions, but also improves the discrimination, which helps improve SLID performance.

Effect of Model Size

As discussed in LSAC-I, one would expect to improve the classifier by having multiple centroids. Let us examine how the number of centroid vectors M affects the performance of LSAC-I. Applying k -nearest-neighboring rule, k is empirically set to 3 in this experiment. In Table 2, it is not surprising to find that performance improves as M increases. In the SVMC, increasing the training corpus will only incur more computation during training. However, in LSAC-I, $L' = L \times M$ exhaustive comparisons need to take place in each test trial. Therefore, it is not practical to have large M . It is interesting to compare the SVMC and LSAC-I that are trained on the same amount of training data. In Table 2, the error rates for SVMC are extracted from Figure 2 for easy comparison.

To reduce computation, LSAC-II uses M mixtures to represent the phonotactic space. With the smoothing effect of the finite mixture model, we expect to use less computation to achieve performance similar to that of LSAC-I. In the experiment reported in Table 3, we find that LSAC-II ($M=1,024$) achieves 14.9% error rate, which is close to the best result in the LSAC-I experiment ($M=12,000$) with much less computation. The results are also illustrated in Figure 4 with the dotted line indicating the best result in the literature [16]. It is noteworthy that when $M > 128$, LSAC-II outperforms the best reported result in the literature.

Table 2. Effect of training data size in LSAC-I & SVMC

#M	1,000	2,000	6,000	12,000
LSAC-I Error (%)	19.8	16.5	15.2	14.8
SVMC Error (%)	18.2	16.2	14.4	13.9

Table 3. Effect of number of mixtures (LSAC-II)

#M	4	16	64	256	1,024
Error (%)	29.6	26.4	19.7	16.0	14.9

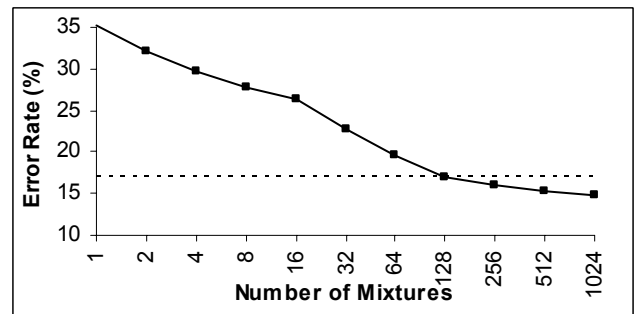


Figure 4. Effect of M (LSAC-II)

From Table 2 & 3, it is found that SVMC outperforms two LSAC classifiers slightly, reporting 13.9% error rate that presents 18.2% error reduction over one of the best reported result, 17.0% [16], on the same test set. This is probably due to the facts that 1) SVMC takes advantages of positive and negative samples during

classifier training resulting in better discriminability; 2) The feature selection is more effective for SVMC than LSAC.

LSAC vs. PRLM

SLID technology has gone through many years of evolution. Many results have been published in the literature on the 1996 NIST LRE database, which provides a good reference point for new technology development. In Table 4, we report the error rates comparison across different models on this test set. It is shown that SVMC & LSAC-II reduce error rates as compared with the best reported results on the 1996 NIST LRE.

Table 4. Benchmark of different models

	PRLM ⁴	P-PRLM & Score Fusion ⁴	LSAC-II	SVMC
Error (%)	22.0	17.0	14.9	13.9

5. CONCLUSION

This paper contributes several new methods to the automatic spoken document classification (SDC) task. First, we propose a non-lexical approach to spoken document tokenization by using a vocabulary of acoustic words; Second, we further explore several feature selection strategies for acoustic vocabulary construction using data driven approach. Third, we propose a *phonotactic-semantic* paradigm to model local phonotactics in an *acoustic word* (AW) and global phonotactics in an *bag-of-sounds* vector, and represent both lexical constraints and latent semantics present in a spoken document. Finally, we formulate two classifiers under two different design criteria. We conclude that the SVMC works well for the feature selection in deciding the AW vocabulary, while the LSAC, formulated by well-established *EM* algorithm, presents good SDC performance. The *phonotactic-semantic* approach not only represents a paradigm shift in spoken language identification (SLID), but also demonstrates 18.2% reduction in error over the benchmark performance on the 1996 NIST LRE data. The results are very encouraging.

In the future, we would like to extend this approach to other spoken document classification tasks. In monolingual SDC, we suggest that the semantic domain be characterized by latent semantics of AWs. Thus it is straightforward to extend the proposed framework to SDC tasks in general. A research over TDT database⁵ is being carried out.

6. ACKNOWLEDGMENTS

The authors are grateful to Dr. Alvin F. Martin of the NIST Speech Group for his advice when preparing the 1996 NIST LRE experiments, to Dr G.M. White and Ms Chen of Institute for Infocomm Research for insightful discussions.

7. REFERENCES

[1] Alshawi, H. Effective utterance classification with unsupervised phonotactic models. *In Proceedings of HLT-NAACL, Edmonton, 2003*, 1-7.

[2] Bellegarda, J.R. Exploiting latent semantic information in statistical language modeling, *In Proc. of the IEEE*, 88, 8 (Aug. 2000), 1279-1296.

[3] Cavnar, W.B., and Trenkle, J.M. N-Gram-Based Text Categorization, *In Proc. of 3rd Annual Symposium on Document Analysis and Information Retrieval*, 1994, 161-169.

[4] Chu-Carroll, J., and Carpenter, B. Vector-based Natural Language Call Routing, *Computational Linguistics*, 25,3 (Sept. 1999), 361-388.

[5] Dai, P., Iurgel, U., and Rigoll, G. A novel feature combination approach for spoken document classification with support vector machines, *Multimedia Information Retrieval Workshop 2003*, Toronto, Canada, Aug 2003.

[6] Duda, R.O., and Hart, P.E. *Pattern Classification and scene analysis*. John Wiley & Sons, 1973.

[7] Garofolo, J.S., Auzanne, C.G.P., and Voorhees, E.M. The TREC spoken document retrieval track: A success story. *In Proceedings of the RIAO 2000 Conference: Context-based Multimedia Information Access*, Paris 2000, 1-20.

[8] Hieronymus, J.L. ASCII Phonetic Symbols for the World's Languages: Worldbet. *Technical Report AT&T Bell Labs*, 1994.

[9] Ma, B., Li, H., and Lee, C.H. An Acoustic Segment Modeling Approach to Automatic Language Identification, *submitted to Interspeech 2005*.

[10] Mladenic, D., Brank, J., Grobelnik, M., and Milic-Frayling, N. Feature selection using linear classifier weights: Interaction with classification with classification models, *SIGIR '04*, Sheffield, UK, 2004, 234-241

[11] Muller, K.R., Mika, S., Ratsch, G., Tsuda, K. and Scholkopf, B. An introduction to kernel-based learning algorithm, *IEEE Trans on Neural Networks*, 12, 2 (Mar 2001), 181-202.

[12] Ng, C., Wilkinson, R., and Zobel, J. Experiments in Spoken Document Retrieval using Phoneme N-gram, *Speech Communication*, 32 (2000), 61-77.

[13] Ng, K., Zue, V.W. Subword unit representations for spoken document retrieval, *In Proc. of Eurospeech 1997*, Rhodes, Greece, 1607-1610.

[14] Salton, G. *The SMART Retrieval System*. Prentice-Hall, Englewood Cliffs, NJ, 1971.

[15] Singer, E., Torres-Carrasquillo, P.A., Gleason, T.P., Campbell W.M., and Reynolds, D.A. Acoustic, Phonetic and Discriminative Approaches to Automatic language recognition, *In Proc. of Eurospeech*, 2003.

[16] Torres-Carrasquillo, P.A., Reynolds, D.A., and Deller, Jr., J.R. Language identification using Gaussian Mixture model tokenization. *In Proc. of ICASSP*, 2002.

[17] Zipf, G.K. *Human Behavior and the Principal of Least effort, an introduction to human ecology*. Addison-Wesley, Reading, Mass, 1949.

[18] Zissman, M.A. Comparison of four approaches to automatic language identification of telephone speech, *IEEE Trans. on Speech and Audio Processing*, 4, 1 (Jan. 1996), 31-44.

⁴ Results extracted from reference [16]

⁵ <http://www ldc.upenn.edu/Catalog/byType.jsp#speech>